Information and Communication Technologies

EPIWORK

Developing the Framework for an Epidemic Forecast Infrastructure

http://www.epiwork.eu

Project no. 231807

D 3.5 Epidemic Data Ontology

Period covered:
Start date of project: February 1st, 2009
Due date of deliverable: January 31st, 2012
Distribution: Public

Date of preparation: January 31st, 2012
Duration:
Actual submission date: January 31st, 2012
Status:

Project Coordinator: Alessandro Vespignani
Project Coordinator Organisation Name: ISI Foundation
Lead contractor for this deliverable: FFCUL

Work package participants

The following partners have taken active part in the work leading to the elaboration of this document, even if they might not have directly contributed writing parts of this document:

João D. Ferreira, Catia Pesquita, Francisco Couto, Mário J. Silva.

Change log

Version	Date	Amended by	Changes
0.1	2012-01-15	Mário J. Silva	First draft.
1.0	2012-01-31	Mário J. Silva	Publication

Epiwork Deliverable 3.5: Epidemic Data Ontology

João D. Ferreira[1], Catia Pesquita[1], Francisco Couto[1], Mário J. Silva[1]

[1]University of Lisbon, Faculty of Sciences, LASIGE, Portugal

31 January 2012

Abstract

We detail our work on evaluating and selecting a network of related ontologies for characterising information relevant to the epidemiological domain. Instead of defining a new ontology from scratch, we propose a Network of Epidemiology-Related Ontologies (NERO), which can be combined to form the core of semantic models in epidemic forecasting infrastructures.

Departing from the metadata model of Epiwork's Epidemic Marketplace (EM), we evaluated existing proposals of ontologies for the epidemiology domain together with other ontologies that, despite having a different purpose, characterise information frequently manipulated by epidemiologists and public health scientists. For the most part, ontologies in NERO are current candidates to the Open Biological and Biomedical Ontologies (OBO) initiative, a large community effort to establish a suite of reference ontologies for the biomedical field.

As part of the integration of Semantic Web technologies into the epidemiological domain, we expose how NERO ontologies can be explored to bring the EM into a more knowledge-oriented repository rather than a simple content database. As such, we delineate the principles through which semantic similarity, ontology matching, and ontology extension can be used to further enhance text mining and other data processing activities in this infrastructure.

Keywords: Epidemic Marketplace, epidemiology-related ontologies, annotation, metadata

Table of Contents

Introduction

Epidemiology research is a truly multidisciplinary subject in the sense that it relies on diverse areas of knowledge, such as biology, medicine, statistics, social sciences and geography. As a scientific field, it requires computational methods to predict the spread of a disease, realistic large scale models, automatic data-collection techniques and, in the context of this deliverable, the creation of platforms for epidemic research and data sharing between research communities and health authorities. Only a framework able to accommodate these methodologies can ultimately deal with all aspects of epidemiology.

Consider as an example a workflow in the context of a hypothetical research project in epidemiology. In that project, some epidemiologists are building a model for death caused by influenza and they need to know the number of deaths caused by this disease over time to fit the parameters of their model. They are interested in building a model that works in France. By means of an appropriate query to an appropriate search engine, they can try to find datasets about "influenza in France". To enable such queries to effectively find the needed resources, the datasets must be correctly annotated so that the information they contain is machine understandable.

If the search locates a matching resource, the epidemiologists can verify that it contains the data they want and eventually use it on their work. However, to fully assist in the retrieval of the relevant information, there must be a suite of auxiliary functionalities that help the epidemiologists in the case of no resource satisfying the query being identified. One of these functionalities is searching for similar resources: a dataset with information about *influenza* in *Europe* is similar to the requested information; likewise, a dataset with information about the symptoms *fever*, *cough*, *headache* and *body aches* in *France* is also relevant, since those symptoms are all associated with influenza. This semantic query expansion results in broadening the scope of the query, which enables the return of relevant resources to the researchers.

Considering now that the researchers want to study the effects of treatment on this disease, they append "*neuraminidase inhibitor*" (a class of antiviral drugs targeted at the influenza virus) to the query. Another possible functionality is inference, which enables the retrieval of resources annotated not only with that term but also with "*oseltamivir*", an example of a neuraminidase inhibitor.

In any case, if these researchers do not find a particularly useful resource on the repository, they can convert their search query to a *request*, detailing what information they need. The request would be automatically annotated with the terms used on the queries (in this

case, both the disease and the location). Suppose now that the researchers fill a request and someone with access to these data finds the request, for instance the authors of a paper about deaths by influenza on USA, France and Australia [1]. Being in possession of the requested data, they can upload and share it with other researchers.

The authors of the paper can also upload this content spontaneously. In this case, there will be no way to infer annotations. To correctly annotate the resource, the authors are presented a form where they can insert the details describing the dataset. Additionally, the content can be automatically analysed to find key terms that can be suggested to the uploaders for annotation. For example, an automatic analyser could find the terms "Influenza" and "France" in the uploaded file, and suggest them back to the authors, who would only need to verify their relevance and accept, correct (in case of a mistake) or reject them.

To enable the functionalities described in the above scenario, it is important to implement a platform for epidemic research that enables data sharing with a semantic basis. The Epidemic Marketplace (EM) is one such a platform. With the establishment of the metadata model for annotating resources submitted to the EM, described in Epiwork Deliverable 3.1 [2], we now face the issue of providing a simple, comprehensive and powerful resource annotation procedure to EM users, which not only increases the functionality of EM but further enhances the consistency of the annotations. By doing so, we will also be contributing to the advance of semantic analysis on epidemiological resources, thereby creating tools to serve epidemic modellers.

To properly introduce these functionalities into the EM, our approach is to identify ontologies that are capable of expressing epidemiologically relevant concepts, such as diseases, which will then serve as a source of annotation terms to the epidemiological resources. This approach has the extra benefit of increasing interoperability with other external services.

In addition, by restricting the annotations to concepts defined in ontologies, we move one step closer to the idea of a Web of Knowledge instead of a Web of Text [3]. These ontologies should be expressive enough to allow their users to faithfully express the contents of the resources and yet strict enough that they allow the full spectrum of Semantic Web tools to operate on them. By doing so, it becomes possible, for example, to perform simple but powerful queries on the EM, or to draw inferences based on the semantics of these annotations [4]. See section "Basic concepts" below for a more detailed description of ontologies and general semantic web technologies.

The main purpose of this deliverable is therefore two-fold:

1. to establish a collection of ontologies that can effectively be used in lieu of a single ontology of epidemiology, namely as a source of concepts that are to be used as annotations to epidemiological resources;
2. to expose how these ontologies can be used to improve usability of the EM.

In particular, we present a list of requirements that an ontology should satisfy to be included in the collection, assessing its usefulness as a new source of concepts for annotating epidemiological resources. This collection, or network, of ontologies will then be used by the EM to enhance the integration and communication of the knowledge it contains among epidemiologists. This resulted in the formulation of a Network of Epidemiology-Related Ontologies (NERO).

Once NERO becomes available, several tools and systems could be developed leveraging it. A possible follow-up step is developing a system capable of performing semantic analysis over the annotations in the EM, improving information retrieval and extraction tasks. Here, we expound how such a system should be implemented, following a modular approach, where each module is responsible for handling a different aspect of this analysis. One of these modules should be able to retrieve related resources by using semantic similarity measures, which can be adapted to different scenarios and particularly to users with different backgrounds. Another module should support the semi-automatic annotation of datasets as they are uploaded: by analysing their content, the system would provide to the uploader a set of suggested ontology terms for dataset annotation. Coupled with user feedback, this semi-automatic process could be improved, potentially to the point where the user is able to suggest concepts not currently present in the network of ontologies. These suggestions could be used by an ontology extension module to improve the ontologies in the network. An additional module should be dedicated to ontology matching to handle the inclusion of new ontologies to the network, merging equivalent concepts and linking related concepts to maintain the coherence in the network and the consistency of the semantic annotations.

Methodology

We started by analysing the current version of the EM metadata model, firstly established in the Epiwork Deliverable 3.1 and further improved in Deliverable 3.4, with particular emphasis on the identification of the domains that should be covered by NERO concepts. This means that NERO is tailored based on the current needs of the EM. The image in **Figure 1** represents this relation between the EM and NERO, showing that the semantic data

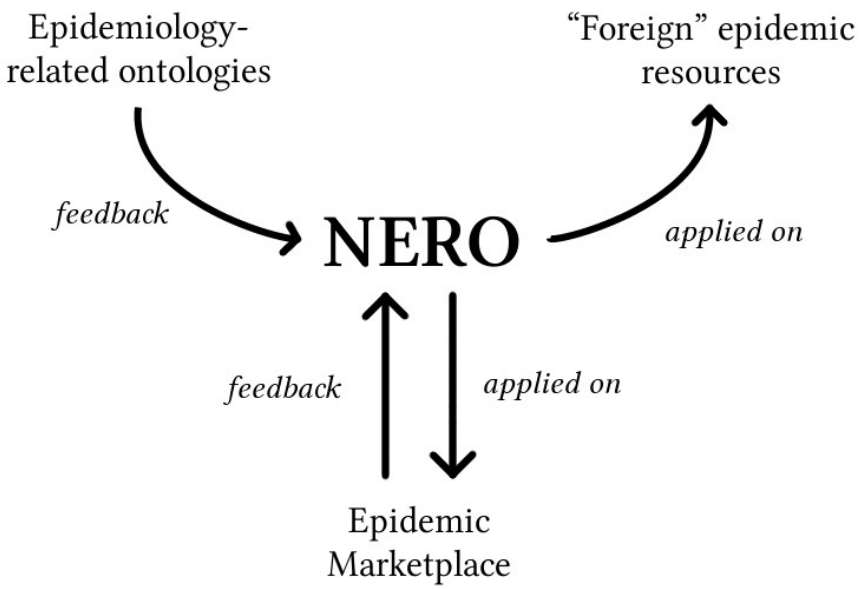

Figure 1. NERO and the EM are related to one another since NERO was created based on specific needs for the EM. However, as illustrated, NERO includes independently-developed ontologies and can be used outside the context of the EM by any epidemiologist, making it a useful resource to the epidemiological research community. In particular, it means that, despite being constructed to work with the EM, NERO has a broader use.

functionalities to be implemented in the EM tightly depend on NERO, but that NERO can serve that same purpose in independently developed systems and tools for epidemic data management.

After this, we discussed which requirements should be fulfilled by the ontologies incorporated in NERO in order to create a concise, comprehensive and good quality network of ontological concepts for the epidemiological domain.

After setting the requirements for the NERO ontologies, we surveyed the existing ontologies that could fulfil them. We began by assessing the quality of currently existing ontologies specific for the epidemics domain, but found that they fail in satisfying many of the requirements. We then considered general-purpose ontologies like UMLS [5] and ontologies with a focussed domain, like the Disease Ontology [6]. After selecting from the identified ontologies those best suited for integration in NERO, we concluded the survey defining a correspondence between these ontologies and the EM metadata elements that they relate to.

This deliverable is organised in three major parts, following this methodology. In the first part, we establish the list of requirements for the NERO ontologies. In the second part, we survey the state-of-the-art in ontologies for the epidemiological domain and establish their correspondence with the EM metadata model. In the final part we explore the capabilities of NERO in an epidemiological scenario and discuss procedures that enable the exploration of the data in epidemiological resources using Semantic Web techniques.

Basic Concepts

Before discussing the results of this deliverable, it is essential to introduce some of the concepts used throughout the document, particularly the concepts of Ontology and Semantic Web technologies.

From an epidemiological point of view, an *ontology* can be seen as a kind of controlled vocabulary extended with relations among the terms in that vocabulary (commonly called *concepts*). Ontologies are one of the ground technologies for the Semantic Web vision. For instance, they allow unambiguous concepts to be reused by different people, thus enabling easy sharing of knowledge between research groups, and as such are a natural candidate for annotating epidemic resources.

Relationships in an ontology can vary from simple subclass-superclass links (for example, *eye* is a subclass of *sense organ*, which is a subclass of *organ,* etc.) to other more expressive links such as *occurs in*, *part of* or *develops from*. It is this set of relations between the concepts that effectively imparts a machine-readable meaning into the concepts. In a nutshell, they are a framework for representing knowledge in a formal manner. For example, an ontology can contain the information that the infection from *Campylobacter* (a genus that contaminates poultry) can result in symptoms like *diarrhoea* and *abdominal pain* [7], enabling the retrieval of resources about abdominal pain with "symptoms caused by individuals of the *Campylobacter* genus" as a search query.

Giving machines a way of processing knowledge through a formal representation (in opposition to processing numbers or text), we are in fact endowing them with the ability to make inference, create proofs, and manage facts in an automatic fashion. One of the most important advantages of using ontologies is the possibility of making a computer understand that, for example, both the concepts *influenza* and *AIDS* refer to infectious diseases.

In a classic approach, if a user presents to a database a query to find resources with the phrase "infectious disease", the database returns a list of resources containing that phrase. However, the query terms "influenza" or "AIDS" also refer to infectious diseases, but they will not be considered by these methods. In fact, there is a need to apply Semantic Web technolo-

gies over these terms to understand that they are related. Inference can be used to close this gap, since it becomes possible to *infer* that AIDS is an infectious disease. It also opens up the possibility of making more complex queries, such as "diseases with a symptom manifested in the lungs".

Furthermore, by using the relations of an ontology, it becomes possible to estimate the *semantic similarity* between two concepts. This technique can be used to improve information retrieval, since it allows the ranking of resources based on their similarity to the actual query. Using the *Campylobacter* example again, the same query could return resources related to "symptoms caused by *Arcobacter*" (a bacteria genus that is closely related to *Campylobacter*).

Requirements of the Network of Epidemiology-Related Ontologies

Instead of building an ontology for the epidemiology domain from scratch, we propose the Network of Epidemiology-Related Ontologies (NERO). The EM and epidemiology in general benefit from this approach because, by reusing well-established ontologies, we rely on the research groups in charge of those ontologies to maintain and curate the concepts that are epidemiologically interesting, thus freeing us from that burden. Moreover, by reusing existing ontologies, we increase the interoperability of the EM with other epidemiological and/or biomedical services, which reduces the effort in aligning ontologies and the resources annotated with their concepts.

In any comprehensive collection of resources with multiple provenances, there must be a set of requirements ensuring and enabling both a good interoperability among those resources and an overall cohesive structure. This section delineates the requirements for NERO. Besides the specific requirements derived from the particular goals of NERO, some of the requirements include adaptations of currently existing principles:

1. principles of the W3C Semantic Web [8];
2. principles of the OBO Foundry [9], a self-appointed foundry responsible for defining standard ontologies in the biomedical domain. See the section "Sources of concepts for NERO" for more details on their work.

This list should be considered as a set of guidelines to work towards an ideal scenario, since it is not expectable that we will find ontologies that completely fulfil them in all relevant areas. In fact, the requirements defined below seem to be sufficient to ensure three important properties:

1. good interoperability between the ontologies;
2. high levels of expressibility in the context of epidemiology;
3. simple, yet powerful, implementation of Semantic Web technologies.

There are ten requirements:

RQ 01: Relevant domain – The most important requirement for incorporating an ontology in NERO is that it should encode a domain of knowledge that is interesting from the point of view of epidemiology. The majority of the concepts of the ontology must be relevant as annotation terms for epidemiological resources. Likewise, the full network should cover almost all of the epidemiological domain, and as such should contain concepts relevant to all the metadata elements of epidemiological resources (diseases, modes of transmission, geographical locations etc.).

RQ 02: Appropriate granularity – To achieve a high coverage of epidemiological concepts, thereby improving the semantic characterization of epidemiological resources, an ontology must provide an adequately detailed representation of the domain. Biomedical and geospatial ontologies tend to comply to this requirement quite well, and in fact some contain many thousands or even tens of thousands of concepts structured over many levels of depth, allowing specific annotations such as the specification of the exact strain of a virus instead of its family. In contrast, the best ontology to describe a given domain can be too granular for the purpose of epidemiological annotation. An example of this is the concept of *photon* in the ChEBI ontology for chemical compounds [10], which is not epidemiologically relevant. In these cases, we should disregard the unwanted branches of the ontology.

RQ 03: Expressiveness with tractability – A wide choice of annotation material is a very important advantage, but being able to manage the ontologies is also a technical requirement, and as such the ontologies must be well structured and tractable from a computational point of view. Specifically, this means that the relationship types must be formally defined and that this definition should be adjusted to the domain in question. For example, in an ontology of anatomy, it does not make sense to have a single subclass-superclass relationship type: relations like *part of*, *arterial supply* and *innervated by* are equally relevant in this domain.

RQ 04: Cross-references to other ontologies – Different domains are usually described in separate ontologies. Sometimes, however, these ontologies are related to one another. For example, an ontology for symptoms and an ontology for diseases model different domains of knowledge, but symptoms are usually associated with diseases and vice-versa. External references that cross from one ontology to another are important in the Semantic Web, since they link together concepts that may not be from the same domain but which share a relation that can be explored. Therefore, an ontology that explicitly stores these cross references has an advantage over one that does not.

RQ 05: Textual definition of the concepts – Given that the concepts of the ontologies in NERO are envisioned to be employed for epidemiological resource annotation, it is very important that its users understand the meaning of the concepts themselves. While the ontology unambiguously encodes the definition of its concepts, it does so in machine-readable code, which is not user-friendly. To offer users the ability to correctly identify the concept they want, the concepts should be complemented with textual definitions as faithful as possible to their ontological meaning.

RQ 06: List of synonyms – Since synonyms are abundant in natural language, particularly in the biomedical field, it is important that the ontology explicitly states these synonyms. For example, when *AIDS* was previously described as part of the Disease Ontology, the concept that we were referring to is actually named "acquired immune deficiency syndrome". Most probably, however, the users expect that "AIDS" refers to the same concept. By stating that the acronym is in fact a synonym to the disease, the ontology becomes more user-friendly.

RQ 07: Popularity – The ontology concepts should be well known in the community of epidemiology, since one of the primary aims of NERO is to be used as a source of concepts for annotation of epidemiological resources. If users are familiar with an ontology, they can more easily choose the correct concept. Furthermore, if an ontology is popular, there is a higher probability that its development does not stall in the foreseeable future, which ensures that NERO ontologies are kept updated with the most current knowledge in the respective domains.

RQ 08: Publicly available – It is generally best to adopt an open-source ontology rather than one that needs licensing or other form of control over usage. This lowers costs, since there is no need to keep an updated license to use the ontology. More importantly, by remaining publicly available, the ontology can be constantly revised by its users, who can submit corrections, suggestions and other improvements to the ontology. Another advantage is that it becomes easier to make suggestions for improvement to the ontology curators based on user needs.

RQ 09: URI persistent identifiers – One of the problems of annotating with ontological concepts is that ontologies are constantly changing in response to the advances in the field, errors found and other factors. This leads to some concepts changing their formal definition, which could ultimately resolve in several annotations becoming wrong. A way of mitigating this effect is giving each concept an identifier that is not semantically relevant and which is never removed from the ontology. Thus, instead of referring to *antiviral treatment* by label, one can use a World-Wide Web dereferenceable identifier, like: *http://purl.obolibrary.org/obo/flu/dev/flu.owl#FLU_0001009*. If this concept ever changes in a way that makes previous annotations incorrect, the term could simply be made obsolete, a new one created, and a link established between the two. The previous annotations would now refer to obsolete terms, but by referring to the link, each annotation can be reviewed and either translated to the new term or changed to the correct one.

RQ 10: Distributed access to the ontology – Several languages have been developed to encode ontologies, most notably OWL (the *de facto* standard in computer science) and OBO format (standard for biomedical ontologies), but other formats exist, from simple tree-like structures described in a text document to tables on a database. Instead of having to cope with all these differences, we require that ontologies be easily accessible through "the cloud", through web services or equivalent, thus enabling a distributed architecture. Moreover, by not having to maintain a local copy of the ontology, there is no need to take special actions in order to keep it up-to-date.

The first three requirements (***domain, granularity and expressiveness***) are scope-related, meaning that they refer specifically to the knowledge encoded in the ontology itself. The others are properties that simplify the tractability of an ontology and improve its usefulness as a scientifically sound source of concepts for annotating epidemiological resources, while ensuring a certain degree of user-friendliness, which is very important in the current context given the aim of NERO.

Additionally, all the above requirements are in accordance to the general EM requirement (defined in Deliverable 3.1). Specifically, they support the sharing and management of epidemiological datasets, the establishment of a community for epidemiological research, the interoperability between this and other modules of the Epiwork project and the use and development of open source solutions.

Sources of concepts for NERO

This section presents (i) a survey on the state-of-the-art in ontological representation of the epidemiological domain, (ii) a list of surveyed ontologies that, despite having been created for other purposes, can be used to describe concepts relevant to this field of research, such as diseases, modes of transmission, demographics or geography, and (iii) a mapping between elements of the Epidemic Marketplace metadata model (refer to Deliverable 3.4) and the ontologies that can be used to fill them. NERO is then defined as the network of these ontologies, alongside with the requirements listed in the previous section, which these ontologies must fulfil. A summary of the considered ontologies and the domain they represent is given in **Table 1**.

Many of the ontologies presented here are already part of an effort to maintain good interoperability and orthogonality between them – the Open Biological and Biomedical Ontologies project (OBO) [9]. OBO includes a variety of biomedical ontologies, some of which are very relevant to the epidemiological domain.

However, we could not find ontologies of good quality for some topics. In such cases, we propose to fill the gap with controlled vocabularies, despite the fact that they are not structured in any ontological sense. Nevertheless, it is expectable that new relevant ontologies will be developed, and given the complementary nature of the ontologies that we incorporate in NERO, these would be able to easily replace the lower quality resources that have been selected.

Ontologies specific to the epidemiological domain

A search, as extensive and exhaustive as possible, was performed on the state-of-the-art concerning the use of ontologies in epidemiology. There have been two attempts at organising epidemiological terminologies in a hierarchical manner [26,27]. These two works describe the ontologies, but neither points the reader to a place where such ontologies can be downloaded or at least browsed.

There have also been a number of automatic systems designed to monitor epidemic surges. One such example is the BioCaster Global Health Monitor [11], an automatic news filter created with the aim of providing "an early warning monitoring station for epidemic and environmental diseases". This system is based on an ontology created by their developers and published in the OWL format, allowing its easy integration with current Semantic Web technologies. It contains almost 2,000 entities. This number may possibly be appropriate for BioCaster purposes (text mining of news articles), but as a source of annotations to epidemic

Table 1. The ontologies found in the survey on the state-of-the-art on epidemiological ontologies, the domain of knowledge they encode and a small comment describing them.

Terminology	Domain	Comment
BioCaster	Epidemiology	ontology used by an automatic news filter to provide early warnings for epidemic diseases [11]
Epidemiology Ontology	Epidemiology	thesaurus of epidemiology developed by the Human Genome Epidemiology Network (HuGENet) [12,13]
Dictionary of Epidemiology	Epidemiology	detailed list of concepts important in the epidemiological field with extensive definitions and usage [14]
UMLS	General	collection of ontologies that promote the creation of interoperable biomedical information systems [5]
MeSH	General	provides an index to articles in biomedical sciences [15]
GeoPlanet™	Geography	structured representation of the world geography [16]
GeoNames	Geography	flat list of geographical locations covering all countries on Earth [17]
Geo-Net-PT	Geography	detailed geospatial ontology of Portugal [18]
OBO ontologies:		
ChEBI	Biochemistry	an ontology of molecular entities focussed on "small" chemical compounds [10]
DOID	Diseases	designed to link disparate datasets through disease concepts [6,19]
ENVO	Environment	supports the annotation of the environment of any organism or biological sample [20]
HP	Symptoms	standardized vocabulary of phenotypic abnormalities encountered in human disease [21]
IDO	Diseases	provides coverage of the knowledge in the infectious disease domain [22]
NCBI Taxonomy	Taxonomy	taxonomic classification of living organisms and associated artefacts [23]
NCI Thesaurus	General	medical terminology focussed on cancer but with general applicability in all health-care [24]
SYMP	Symptoms	captures and documents both symptoms and signs in medical literature [19]
TRANS	Disease transmission	describes how a pathogen is transmitted from one host, reservoir, or source to another host [19]
VO	Vaccines	representation of vaccine knowledge [25]

resources, it is very poor. For example, only five countries appear in the ontology, and while there are a number of diseases and syndromes, they are very shallowly organised (diseases are instances of *Avian Disease*, *Human Disease* or other similar classes, all of which are subclasses of the generic concept *Disease*, for a maximum of three levels of depth). Concepts of the therapeutics domain are not well represented (in fact, there is a *therapeutic role* concept, but no other entity in this area) and there is no concept of vaccination. Overall, we observe that the majority of concepts in this ontology is better represented in other ontologies, because they are focussed on a more specific domain.

Another ontology built especially for epidemiological studies is the *Epidemiology Ontology*, developed by the Human Genome Epidemiology Network (HuGENet) [12,13]. This is not as well structured as the BioCaster ontology, as it consists of a single hierarchy of terms related by a single relationship type. This raises situations such as *Person* being a described under *Hypothesis Formulation from Descriptive Studies* or *Hospital* under *Notifiable disease.* This ontology contains 791 distinct concepts, some of which also appear in the *Dictionary of Epidemiology* [14], a dictionary that also contains a very detailed list of concepts important in the epidemiological field. Despite the alphabetic organisation and the absence of a hierarchy, this is a good quality source for epidemiological concepts, with entries for almost 2000 concepts, each with a detailed description of its meaning and some form of structure given in the descriptions as references to other entries. Considering all the domains of NERO, and given the low coverage of the Epidemiology Ontology and the Dictionary of Epidemiology in domains such as geography or diagnostic methods, we believe that they have limitations. Just like BioCaster, however, they can help by providing a sense of which concepts should be modelled in an Epidemiological resource.

Other ontologies containing epidemiological concepts

Given the low suitability of those resources to be incorporated in NERO, we moved our focus to ontologies containing relevant concepts to the epidemiology domain. In this context, the relevant domains were assumed to be the ones in need for the EM metadata model, which agree with the domains of knowledge represented in BioCaster, the Epidemiology ontology and the Dictionary of Epidemiology.

Some research has been conducted based on the *use* of existing ontologies rather than the *development* or *creation* of new ones. Such works are based on ontologies containing epidemiologically relevant concepts, but which were not designed with that specific domain of knowledge in mind. Ontologies in this category include the Unified Medical Language System (UMLS), a collection of ontologies and terminologies that "promote the creation of

more effective and interoperable biomedical information systems and services" [5], and Medical Subject Headings (MeSH), a controlled vocabulary used to index articles in biomedical sciences [15]. These resources can be seen as hierarchies of terms, where a term directly descends from one or more terms, thus creating a graph-like structure that can be easily navigated.

As an example, consider the work of Xu H. *et al.*, which uses UMLS to mine for epidemiologically relevant concepts in texts [28]. While such resources could prove useful, the UMLS is a very large resource, with over one million concepts; properly scanning through this terminology and determining which of these concepts are relevant in an epidemiological sense would be too colossal a task for the typical epidemic modeller.

Additionally, MeSH is relatively unstructured and makes use of a single relation, *narrower than*. For example, *Axial length* and *Eyebrow* are categorised under *Eye*, but one is a *property* and the other is a *nearby* structure. Likewise, *Eye* is both categorised under *Sense Organs* and *Face*, but while it *is* a sense organ, it is *part* of the face. MeSH makes no distinction between these semantic relations, which we consider one of the main drivers for the use of ontologies.

There are other limitations with UMLS and MeSH: since they have a very generic and broad domain, the addition of new concepts is non-trivial, as there is a high risk of introducing errors and inconsistencies. In fact, it is known that UMLS houses many inconsistencies [29]. Finally, these two resources are not published in a standard Semantic Web format, meaning that they do not integrate well with Semantic Web technologies.

In face of these issues, we turned to attempts in the biomedical field to create and organise more formal ontologies. There is one project that should be highlighted: the Open Biomedical and Biological Ontologies (OBO). This is a project run by the OBO Foundry that aims to provide a suite of orthogonal interoperable reference ontologies in the biomedical domain [9]. The OBO Foundry defines a set of principles that must be fulfilled by an ontology before it is included (in fact, some of the requirements of NERO were inspired in OBO principles). There are currently eight OBO ontologies, but 97 other candidates are presently working to fulfil the required principles for being endorsed by the OBO Foundry. Given that OBO's set of principles enforce good quality ontologies by promoting good practices in ontology development, and that any one of these ontologies, both supported and candidate, thrives to fulfil those principles, we included some of them in NERO (see **Table 1** for the list of the considered ontologies), enabling the comprehensive description of biological and biomedical aspects of the resources in the marketplace. Because the ontologies of OBO span over many of the biological and biomedical domains of knowledge, these domains can be well covered in NERO.

Concepts from geography, demographics, etc., which are not biological, must be retrieved from other resources. Yahoo! GeoPlanet™ [16] contains a representation of the world geography, and is in fact a very good candidate for inclusion in NERO. Other geographical ontologies were considered, such as GeoNames [17] and Geo-Net-PT [18]. GeoNames is a flat dictionary of locations on Earth, lacking an ontological structure. For instance, there is no information about the relation between Italy and Rome (its capital) or Italy and France (one of its neighbours). Geo-Net-PT is an ontology of the Portuguese territory and, even though it is rich in detail, it covers a small scope of the Earth. However, there are correspondences between Yahoo! Geoplanet™ and Geo-Net-PT [30]; therefore, if a more detailed annotation is required, Geo-Net-PT would be a good complement in the area it covers.

We have been unable to find ontologies that specifically represent demography or social and economic conditions, and suspect that none exist that are publicly available. As such, we will have to rely on other resources such as UMLS, MeSH, the Epidemiology Ontology or the Dictionary of Epidemiology for those domains. In this context, it is important to mention that there are some databases and tables with this information, such as the demographic data of the United Nations [31]. These databases and tables do not allow the application of Semantic Web technologies, since they lack a machine-readable semantic; therefore, they would need to be introduced and curated in an ontological format (e.g., through triplification [32]) before being included in NERO.

A summary of the survey

UMLS and MeSH cover most of the epidemiological domain in a way that satisfies epidemic modellers' needs. However, they contain many other irrelevant terms, and are complex and not as well maintained as needed for a fully Semantic Web approach. We have also found other resources that try to represent epidemiological concepts: BioCaster and the Epidemiology Ontology. They are not comprehensive enough for the entire epidemiological domain, yet they offer an insight into what an epidemiological network of ontologies should contain and how it should be organised, since they contain the branches of knowledge that are required in epidemiology (diseases, modes of transmission, locations, social conditions, etc.). By crossing the information contained in those resources with the EM metadata model, we have come to the conclusion that ontologies from the OBO project are the most appropriate terminologies for the biomedical portion of the EM, since, together, they span over a large amount of the biological and medical domains of epidemiology.

For geographical information, we incorporated Yahoo! GeoPlanet™ in NERO, based on its better quality versus the other possibilities. For the domains of demography and social and

economical conditions, some of the branches of the general purpose ontology MeSH were considered. Additionally, the Epidemiology Ontology has also been included in order to further increase the coverage of these domains in NERO.

Table 2. Evaluation of the terminologies considered in this section, based on the requirements of NERO. **Legend**: **Y** – terminology fulfils the requirement; **±** – terminology partly fulfils the requirement; **N** – terminology does not fulfil the requirement; **N+** (on requirement 2) – terminology is more granular than required.

	Requirements									
	Relevant domain	**Appropriate granularity**	**Expressive & tractable**	**Cross-references**	**Textual definitions**	**Lists of synonyms**	**Popularity**	**Publicly available**	**Identifiers**	**Distributed access**
Terminology	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
BioCaster	Y	N	±	Y	Y	Y	N	Y	N	N
Epidemiology Ontology	Y	Y	N	N	N	Y	N	Y	N	N
Dictionary of Epidemiology	Y	Y	N	N	Y	±	N	Y	N	N
UMLS	Y	N+	N	Y	Y	Y	Y	±	Y	N
MeSH	Y	N+	N	N	Y	Y	Y	Y	Y	N
GeoPlanet™	Y	Y	N	Y	±	Y	Y	±	Y	Y
GeoNames	Y	Y	N	Y	N	Y	Y	Y	Y	Y
Geo-Net-PT	Y	N	Y	Y	±	Y	N	Y	Y	Y
OBO ontologies:										
ChEBI	Y	N+	Y	Y	Y	Y	Y	Y	Y	Y
DOID	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
ENVO	Y	±	Y	±	Y	Y	N	Y	Y	Y
HP	Y	±	Y	Y	Y	Y	Y	Y	Y	Y
IDO	Y	±	Y	N	Y	N	Y	Y	Y	Y
NCBI Taxonomy	Y	N+	Y	N	N	N	Y	Y	Y	Y
NCI Thesaurus	Y	N+	Y	Y	Y	Y	Y	Y	Y	Y
SYMP	Y	Y	Y	N	N	Y	N	Y	Y	Y
TRANS	Y	N	Y	±	Y	N	N	Y	Y	Y
VO	Y	Y	Y	N	Y	N	N	Y	Y	Y

A more graphical summary is shown on **Table 2**, which details the requirements fulfilled by each of the ontologies found, and **Table 3**, which contains the ontologies incorporated in NERO.

Correspondence between the new model and existing ontologies

Given the discussion on the existing ontologies with relevant domains in the epidemiological field and the list of requirements, we synthesised which of those ontologies were judged fit to be used as a source of annotation terms for each of the elements in the EM metadata model. **Table 3** shows a mapping between the relevant metadata elements of the new model and the ontology or ontologies that are best suited to provide values for them. As an illustration of the use of NERO in the EM, consider the metadata of a hypothetical EM resource in Appendix A.

Table 3. This table maps the metadata elements of the Epidemic Marketplace metadata model into ontologies that contain concepts useful to describe epidemiological resources. Each element can be mapped to more than one ontology, which is useful when neither covers 100% of the domain in question.

Metadata element	Proposed ontologies	Provenance
`<em:diagnosticMethod>`	NCI Thesaurus	OBO candidate
`<em:disease>`	DOID	OBO candidate
	IDO	OBO candidate
`<em:drug>`	ChEBI	OBO Foundry
`<em:symptom>`	SYMP	OBO candidate
	HP	OBO candidate
`<em:host>`	NCBI Taxonomy	OBO candidate
`<em:pathogen>`	NCBI Taxonomy	OBO candidate
`<em:vector>`	NCBI Taxonomy	OBO candidate
`<em:transmission>`	TRANS	OBO candidate
`<em:vaccine>`	VO	OBO candidate
`<em:environment>`	ENVO	OBO candidate
`<em:location>`	GeoPlanet™	Yahoo!
`<em:demography>`	Branches of MeSH	NLM/NIH
	Epidemiology Ontology	HuGENet
`<em:socioEconomicCondition>`	Branches of MeSH	NLM/NIH
	Epidemiology Ontology	HuGENet
`<em:geographicalEncoding>`	??	??

The element `<em:geographicalEncoding>` (used to annotate resources with the type of geographical information included, such as “map”, “coordinates” or “ontology concepts”) is denoted with question marks since no good ontology has been found for it. We intend to harness the knowledge of EM users by providing them with the ability to write a free-text term or to choose from less appropriate terminologies, such as the Dictionary of Epidemiology. These non-ontological terms can be used to complement the ontologies encompassed by NERO, or possibly to adapt the less appropriate terminologies to more semantic-friendly hierarchies.

If necessary, this same behaviour can be applied to the other elements of the metadata model as well. For example, if a user wants to annotate a resource with a diagnostic method that does not exist in the NCI Thesaurus or a mode of transmission absent from TRANS, they should be given the opportunity to express their annotation as free text. By analysing these values, we can therefore extend the ontologies in NERO to better attend to the Epidemic Marketplace needs (see also the section “Ontology Extension” on Appendix B).

Beyond semantic annotation with NERO

Once the EM is populated with annotated resources, it will be possible to exploit these annotations to perform complex semantic analyses on diverse tasks, such as information retrieval and information extraction. These tasks will provide epidemiologists, particularly the epidemiology modellers, with tools that enable an easy discovery of models and the parameters to use in them.

There are two main challenges in accomplishing this goal. The first is to define a way to effectively compare resources that are annotated using different sets of ontologies, i.e. how to compare a resource annotated with HP and ChEBI, to another annotated with ChEBI and NCI Thesaurus. This problem is relevant within the Epidemic Marketplace, where different resources will have different domains, and as such will be annotated using different ontologies. It also affects the general use of NERO, since resources annotated with NERO concepts may be, at some point, compared to resources annotated with other ontologies. The second challenge resides in providing a contingency plan for handling cases where few or no annotations exist, which translates to how to generate annotations in an automated or semi-automated fashion for a given resource. Although we expect this situation to become increasingly less frequent as EM gains momentum, it will always remain a necessity to complement manual annotation.

This system will comprise two main modules, each with an auxiliary module: the first challenge will be addressed by a semantic similarity module coupled to an ontology matching one, while the second challenge will be undertaken by a text mining module backed up by a semi-automated ontology extension one.

The semantic similarity module will address the issue of similarity between resources annotated using multiple ontologies. Since current implementations of suitable semantic similarity measures only span a single ontology, we will need to develop methods that are able to perform comparisons across multiple ontologies, particularly handling non-hierarchical relations. The authors have been actively working in semantic similarity in the past few years, and have made relevant contributions particularly in the field of biomedical ontologies [33-35]. This module improves information retrieval by allowing a user to find resources that are similar to an input resource. For instance, a user can be interested in finding all resources related to viral diseases in children. The system can retrieve resources related to this query by calculating the similarity between it and the annotated resources in the EM. Alternatively, the user can also use as input a given resource and find all related ones, according to different aspects: while a physician may be more interested in finding resources

with similar therapeutics, a biologist may prefer resources with a similar vector. To accommodate these scenarios, the system will allow the assignment of weights to different ontologies, which modulates their contribution to the final similarity.

Semantic similarity across multiple ontologies can exploit correspondences between their concepts, particularly through the use of cross-references. When such resources are unavailable, ontology matching techniques can be used to automatically create them. This will allow the EM to find relations between concepts from different ontologies, increasing the accuracy of similarity and, as such, the performance of information retrieval. We have previous experience in this field, particularly in the areas of biomedical and geographical ontology matching [30,36].

Measures of semantic similarity rely on the resources being annotated with ontology concepts. The text mining module will handle cases where these annotations are not sufficient by extracting relevant and non-trivial information from the content of EM resources, and then creating new annotations. This is particularly relevant in poorly annotated resources, since their usefulness to the community is directly dependent on their being able to be easily retrieved. One of the main purposes of this module is to facilitate the process of annotation to users, since the flexibility of the metadata model does not enforce complete annotation. By analysing the content of the files being uploaded, this module is responsible for mining the text to find, for example, disease concepts or geographical places. These will then be suggested to the users, which can accept or reject them. This will improve not only the quantity but also the quality of annotations, contributing to a better performance in information retrieval and to a more coherent corpus of annotations. Likewise, this is an area where members of this project have expertise [37].

When NERO ontologies do not have a sufficient degree of specificity, new concepts can be added using the semi-automatic ontology extension module, which will be capable of automatically suggesting new concepts and relations. New concept suggestions can be derived from text or external ontologies and resources, or more interestingly from the free text annotations made by EM users.

See Appendix A for a more detailed description of these technologies.

The integration of these modules will result in a full fledged system for Semantic Web based information retrieval and extraction over the resources in the EM that is able to support a variable degree of user involvement.

Conclusions and final remarks

In this deliverable we established a Network of Epidemiology-Related Ontologies (NERO) to be used as a source of annotation of epidemiological resources. The proposal of this network of ontologies results from our experience developing the EM metadata model. In the future, it will support annotation of epidemiological resources as well as the application of Semantic Web technologies over them. In particular, we plan to fully integrate NERO in EM.

A crucial first step to fully realize these features in the EM is the integration of NERO ontologies into the resource upload and edition forms, thus providing EM users with a structured annotation procedure. We expect that this will make the annotation process not only easier but also more complete, since users will have a standard set of concepts to choose from. This will eventually result in a corpus of annotated epidemiological resources, over which the information retrieval and extraction system can operate. The implementation of this system will rely on our team's expertise in information retrieval and extraction and the feedback of other partners to ensure that the system is providing users with high quality results.

By providing better tools to search EM resources, these will become more accessible to EM users, fostering the sharing of epidemic resources. Likewise, this increased accessibility will make it easier to find researchers working on related fields, encouraging a virtual community for epidemic research. In fact, NERO is able to serve all the epidemiology community, since it is not bound to the EM but can subsist on its own. For example, the research teams in charge of other work packages (WP) in the Epiwork project can also benefit from using NERO, particularly those in WP2 and WP5. WP2 is responsible for developing approaches to identify and quantify modularity in spatially structured and heterogeneous meta-populations and contact networks. The geospatial information that NERO encodes can be of great interest here. WP5 is responsible for providing validated data through ICT applications, in particular the internet-based monitoring system InfluenzaNet. Semantically annotating the data collected in this WP is a major step in its analysis, and NERO can serve as the source of concepts for this annotation.

The establishment of this network of ontologies contributes, therefore, to an improvement for all the community, particularly on the topics of sharing and reusing epidemiological resources.

Appendix A – The metadata of a possible EM resource

As an example, we present the metadata of a fictitious Epidemic Marketplace resource of the type *Dataset*. As it will be common to all resources of the marketplace, not all metadata elements are defined. An XML comment was added next to the elements that were filled-in with ontology concepts, to contextualise the choices. Such comments are not needed in the EM, since its URI can, in general, be used to dereference the name of the concepts along with related information retrieved from the cloud.

This example uses the version of the EM metadata model provided in Deliverable 3.4. For an explanation of the elements of the metadata model that use NERO concepts, see **Table A.1**.

Table A.1. This table explains the usage of each one of the metadata elements that should be filled in with NERO concepts.

Metadata element	Definition
`<em:diagnosticMethod>`	Diagnostic test used to obtain data in the resource; or the protocol defined in the resource
`<em:disease>`	Disease that is mentioned in the contents of the resource.
`<em:drug>`	Chemical compounds associated with in the resource, as cause and/or treatment of a disease
`<em:symptom>`	Symptom that is mentioned in the contents of the resource
`<em:host>`	Taxonomical reference to the organism or organisms that are hosts of a disease
`<em:pathogen>`	Taxonomical reference to the organism or organisms that are pathogens in a disease
`<em:vector>`	Taxonomical reference to the organism or group of organisms that are vectors of a disease
`<em:transmission>`	Mode of disease transmission mentioned in the resource
`<em:vaccine>`	Vaccine used in the study the dataset refers to
`<em:environment>`	The environmental data contained in the resource
`<em:location>`	The spatial coverage of the resource
`<em:demography>`	The type of demographic elements contained in the resource
`<em:socioEconomicCondition>`	The type of social and/or economical data contained in the resource
`<em:geographicalEncoding>`	The type of geographic data contained in the resource

```
<em:em xmlns:em="http://epimarketplace.net/namespace/">
   <em:title>Example Epidemic Dataset</em:title>
   <em:identifier>empid:2345</em:identifier>
   <em:generalDescription>
      <em:description>
         This dataset contains some exemplifying data that can be consulted by anyone
         through the Epidemic Marketplace.
      </em:description>
      <em:DOI>doi:1234-567</em:DOI>
      <em:format>application/pdf</em:format>
      <em:format>text/csv</em:format>
      <em:language>en_US</em:language>
      <em:subject>Mobility</em:subject>
      <em:type>http://epimarketplace.net/namespace/dataset</em:type>
      <em:URL>http://example.com/epidemic_dataset</em:URL>
      <em:version>1.0</em:version>
   </em:generalDescription>
   <em:date>2010-10-13</em:date>
   <em:dateSubmitted>2011-11-11T16:37:34Z</em:dateSubmitted>
   <em:organisation>
      <em:organisationName>Example Organisation</em:organisationName>
      <em:organisationURL>http://example.org</em:organisationURL>
   </em:organisation>
   <em:uploader>
      <em:uploaderName>João D Ferreira</em:uploaderName>
      <em:uploaderOrganisation>LaSIGE</em:uploaderOrganisation>
   </em:uploader>
   <em:time>
      <em:from>2010-09-01</em:from>
      <em:to>2010-09-30</em:to>
   </em:time>
   <em:biologogicalInformation>
      <em:disease>
         http://purl.obolibrary.org/obo/DOID_8659<!--chickenpox-->
      </em:disease>
      <em:symptom>
         http://purl.obolibrary.org/obo/SYMP_0000009<!--blister-->
      </em:symptom>
      <em:host>
         http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=9606<!--Homo spiens-->
      </em:host>
      <em:pathogen>
         http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=10338
         <!--Varicella-zoster virus, strain Dumas-->
      </em:pathogen>
      <em:symptom>
         http://purl.obolibrary.org/obo/SYMP_0000090<!--enchephalitis-->
      </em:symptom>
      <em:transmission>
         http://purl.obolibrary.org/obo/TRANS_0000007<!--direct contact-->
      </em:transmission>
      <em:vaccine>
         http://purl.obolibrary.org/obo/VO_0000728<!--chickenpox virus vaccine-->
      </em:vaccine>
   </em:biology>
```

```
   <em:environment>
      http://purl.obolibrary.org/obo/PATO_0000146<!--temperature-->
   </em:environment>
   <em:demography>http://epimarketplace.net/freetext?q=life+expectancy</em:demography>
   <em:socioEconomicCondition>
      http://epimarketplace.net/freetext?q=household+average+income
   </em:socioEconomicCondition>
   <em:location>http://where.yahooapis.com/v1/place/551801<!--Vienna--></em:location>
   <em:location>http://where.yahooapis.com/v1/place/742676<!--Lisbon--></em:location>
   <em:location>http://where.yahooapis.com/v1/place/725003<!--Torino--></em:location>
   <em:bibliographicCitation>
      <em:refCitation>
         Example Organisation (2010). An example dataset for epidemiology. Journal of
         Examples, 2:11, pp. 100-112.
      </em:refCitation>
      <em:refDOI>doi:9876-567</em:refDOI>
   </em:bibliographicCitation>
   <em:rights>
      <em:copyright>Public Domain</em:copyright>
   </em:rights>
</em:em>
```

Appendix B – Semantic Web technologies

Semantic Similarity

Following Pesquita *et al.* [38], we define a semantic similarity measure as a function that returns a numerical value reflecting the closeness in meaning between two ontology concepts or two sets of concepts annotating two resources. Although similarity can sometimes be achieved with more straightforward methods (the alignment of genes and proteins through the BLAST algorithm, or the distance between geospatial locations), many properties, like the function of gene products or the relatedness between geospatial locations, can only be compared in the context of ontologies since they lack other formal representations.

Measures of semantic similarity can be categorized in two groups based on the ontological information they use. Edge-based methods use the relations of the ontology; node-based methods use the concepts themselves. These methods can compute the similarity value based on the ontology information alone (intrinsic methods) or they can also use information that is not encoded in the ontology (extrinsic methods). For instance, the most successful semantic similarity methods applied to GO [33] use the concept of information content (IC). IC measures a concept's specificity independently of its depth in the ontology, since it is based on its frequency of annotation in a corpus [39]. This notion can be exploited by semantic similarity measures to calculate the amount of information two concepts share. This can be achieved by finding the IC of the common ancestry between the concepts, which can be given by the IC of their most informative common ancestor (MICA) or the sum of the ICs of all their disjoint common ancestors (DCA), i.e. the common ancestors that do not subsume any other common ancestor [16].

The true applicability of semantic similarity methods, however, lies in the fact that they enable the comparison of resources that are annotated with concepts of one ontology. These comparisons are either pairwise, where the similarity values between the individual annotations of both resources are combined to produce a single value, or groupwise, where the resources are translated into sets, graphs or vectors and then compared using the appropriate techniques [38].

There are some cases where semantic similarity between two resources cannot be calculated using a single ontology. Epidemiology needs biological information about the disease and the mode of transmission, but also uses geographical information to describe the spread of the disease. Comparing these complex models using a single ontology is like comparing diseases only by their symptoms. For example, chest pain is common to myocardial infarc-

tion and acid reflux, two unrelated diseases, but there are cases of myocardial infarction without chest pain. Thus, using a single ontology is to disregard important information that other ontologies have to offer. In fact, the true usefulness of semantic similarity in the EM lies in its ability to compare resources annotated with various NERO concepts.

Semantic similarity measures developed for the EM need to be able to compare complex resources annotated across multiple ontologies, and as such should be able to:

1. compare concepts from the same ontology;
2. compare concepts from different ontologies, provided these have cross-references or bridges between them;
3. compare sets of concepts from different ontologies using various grouping techniques;
4. define parameters that allow distinct weighting strategies to be applied over the grouping techniques in order to enable similarity scores to reflect users' interests.

Ontology Matching

One way to implement semantic similarity measures for concepts from different ontologies is to rely on correspondences between related concepts, in order to create a common structure. This solution is restricted to ontologies with the same or related domains. Such correspondences already exist for some ontologies that provide cross-reference resources. When these resources are unavailable, ontology matching techniques can be used to automatically create them.

Ontology matching has been defined as “finding correspondences between semantically related entities of different ontologies” [30]. These correspondences may represent not only equivalence, but also other kinds of relations, such as consequence or relatedness. Given the manpower necessary to create these matches manually, there has been an increased interest in matching biomedical ontologies in an automated fashion.

Ontology matching algorithms can exploit ontology internal information, such as concept's properties (labels, synonyms, data types, etc.) and relations, or external knowledge in the form of annotation corpora and other ontologies, resources and alignments. Matchers that focus on comparing pairs of concepts individually, are called element-level matchers, whereas matchers that take in consideration multiple concepts and their relationships are called structural-level matchers. Element-level similarities can be explored by structural-level approaches that use global similarity computation techniques. These techniques assume that the similarity between two concepts depends on the similarity between their adjacent

concepts, and therefore similarities can be propagated throughout the ontologies to provide a final alignment, i.e. the optimized set of matches between two ontologies [41,42].

Given the broad range of domains of NERO, ontology matching strategies need to be able to handle the different challenges they present. For instance, many biomedical ontologies are very large and support few types of relationships, which can hinder their alignment. On the other hand, they contain rich textual information that can be exploited by lexical matching.

Text Mining

Some semantic similarity measures and ontology matching techniques depend on the existence of annotations with ontological concepts.

Text mining generally concerns the process of extracting relevant and non-trivial information and knowledge from unstructured text, usually a collection of documents. After creating a structured representation of texts, text mining systems use a rule-based or a case-based approach. The rule-based approach relies on patterns identified by an expert which contain relevant information. These patterns are then converted to rules to identify the relevant information in the rest of the text. The main bottleneck of this approach is the manual process of creating rules and patterns, which is time-consuming and, in most cases, unable to derive a set of rules that encompass all possibilities. The case-based approach relies on a predefined set of texts annotated by an expert, which is used to learn a model for the rest of the text. The main bottleneck of this approach is the selection of a training set large enough to enable the creation of a model accurate for all texts. None of these knowledge representation techniques subsumes the other: the knowledge enclosed in a rule is normally not fully expressed by a finite set of cases, and it is difficult to identify a set of rules encoding all the knowledge expressed by a set of cases.

Text mining techniques have varying degrees of success, depending on the domain of their application. For instance, while the performance in annotating geographical entities has reached high levels, these results have not been reproduced for the biomedical domain, due to the complexity in describing biologic concepts and entities. In biomedical literature, we can often find synonyms and homonyms. Moreover, common English words are frequently used as names (e.g. fruity and cactus are actually gene names), or as acronyms (e.g. AND, ETC), which makes it difficult to recognize biological entities in text. Bottom line, the information to extract is complex, and therefore it is almost impossible to derive a rule without having a significant number of exceptions.

Recent advances in text mining of biomedical literature already achieved acceptable levels of accuracy in recognising gene and protein names in text. However, the extraction of more

complex biomedical entities and relationships, such as functional annotations, is still far from being solved [43-45].

Additionally, by applying text-mining techniques on the description of the entries in the Dictionary of Epidemiology, we can discover relations between these entries, thus introducing the concept of machine-readable semantics to the dictionary, which will contribute to the amount of knowledge represented in NERO. The next example illustrates this procedure.

In the Dictionary of Epidemiology, *Disease model* is defined as "a quantitative simulation of the natural history of a disease (incidence, progression, prognosis, etc.) based on epidemiological data. A public health model is population-based and is used in planning and evaluating health services, whereas a clinical model is used in individual patient care." From this we can extract this information:

- *Disease model* is a *Quantitative simulation*
- *Disease model* is based on *epidemiological data*
- *Public health model* is a *Disease model*
- *Public health model* is a *Population-based model*
- *Public health model* is used in *Planning and evaluation of health services*
- *Clinical model* is a *Disease model*
- *Clinical model* is used in *Individual patient care*

Ontology Extension

When ontologies do not have a sufficient degree of specificity, new concepts can be added through a process called *ontology extension*. This is particularly relevant in the life sciences domain, where knowledge is complex and continuously changing and growing. These ontologies can never be considered complete, but always have to adapt to the new understanding of biomedical knowledge through an iterative process [46]. One solution that eases the burden of adapting the ontology to new knowledge is the application of semi-automated ontology extension techniques, which are capable of automatically suggesting new concepts and relations to add to the ontology. These are usually adapted from related areas of ontology engineering: ontology learning and ontology matching.

The most common data source used in ontology extension is natural language text, due to its availability and coverage. Therefore the majority of ontology extension techniques are based on the extraction of terms from text using term relevance measures [47,48]. The insertion of these new concepts at the appropriate position in the ontology is usually addressed by machine learning techniques that classify the new concept into an existing ontology class, or

are based on co-occurrence patterns and syntactic patterns [49,50]. However, other resources such as related ontologies, can also be used. More interestingly, free text annotations made by EM users can also be a relevant source of new concepts.

References

[1] Viboud C, Boëlle P-Y, Pakdaman K *et al.* (2004). Influenza Epidemics in the United States, France, and Australia, 1972-1997. *Emerging Infectious Diseases*, **10**(1):32-39.

[2] Lopes LF, Silva F, Couto F, Silva M (2009). Epiwork Deliverable D3.1: Meta-model – Initial Specification, Catalogue of Relevant Data, Platform Requirements.

[3] Berners-Lee T, Hendler J (2001). Scientific publishing on the semantic web. *Nature*, **410**:1023-1024.

[4] Linked Data. http://www.w3.org/standards/semanticweb/data Accessed on 2011-12-16.

[5] Unified Medical Language System. http://www.nlm.nih.gov/research/umls/ Accessed on 2011-12-16.

[6] Osborne J, Flatow J, Holko M et al. (2009). Annotating the human genome with Disease Ontology. *BMC genomics*, **10**(Suppl 10):S6.

[7] Skirrow MB (1977). Campylobacter enteritis: a "new" disease. *British medical journal*, **2**(6078):9-11.

[8] Koivunen MR, Miller E (2001). W3C Semantic Web Activity. *Semantic Web KickOff in Finland*. HIIT Publications. http://www.w3.org/2001/sw/ Accessed on 2011-12-22.

[9] Smith B, Ashburner M, Rosse C *et al.* (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology* **25**:1251-1255.

[10] de Matos P, Alcántara R, Dekker A, *et al.* (2010). Chemical entities of biological interest: an update. *Nucleic Acids Research*, **38**(suppl 1): D249-D254.

[11] Collier N, Doan S, Kawazoe A *et al.* (2008). BioCaster: detecting public health rumors with a Web-based text mining system. *Bioinformatics*, **24**(24):2940-2941. doi:10.1093/bioinformatics/btn534.

[12] Khoury MJ, Dorman JS (1998). The human genome epidemiology network. *American journal of epidemiology*, **1**:1-3. http://www.hugenet.org.uk/ Accessed on 2011-12-16.

[13] Guidelines for the Epidemiological Ontology. http://www.hugenet.org.uk/resources/informatics/Ontology_Version_1.pdf Accessed on 2011-12-16.

[14] Porta M. (2008). *A Dictionary of Epidemiology*. Oxford University Press, USA.

[15] National Library of Medicine (2000). *Medical subject headings: main headings, subheadings, and cross references used in the Index Medicus and the National Library of Medicine Catalog.* 1st ed. Washington, DC: U.S. Department of Health, Education, and Welfare.

[16] Yahoo! GeoPlanet™. http://developer.yahoo.com/geo/geoplanet/ Accessed on 2011-12-16.

[17] GeoNames. http://www.geonames.org/ Accessed on 2012-01-09.

[18] Lopez-Pellicer FJ, Chaves M, Rodrigues C, Silva MJ (2009). Geographic Ontologies Production in Grease-II. Technical Report. TR 09-18. Universidade de Lisboa, Faculdade de Ciências, LASIGE, doi:10455/3256.

[19] Schriml LM, Arze C, Nadendla S *et al.* (2010). GeMInA, Genomic Metadata for Infectious Agents, a geospatial surveillance pathogen database. *Nucleic Acids Research*, **38**(Suppl 1):D754.

[20] The EnvO Project. http://gensc.org/gc_wiki/index.php/EnvO_Project Accessed on 2012-01-05.

[21] Robinson PN, Mundlos S (2010). The human phenotype ontology. *Clinical genetics*, **77**(6):525-534.

[22] Cowel LG, Smith B (2010). Chapter 19 "Infectious Disease Ontology", *in* Sintchenko V. (Editor) *Infectious Disease Informatics*, pp. 373-395.

[23] Sayers EW, Barrett T, Benson DA *et al.* (2008). Database resources of the National Center for Biotechnology Information. *Nucleic Acid Research*, **37**(Database issue):D5-15.

[24] Sioutos N, Coronado S, Haber MW *et al.* (2007). NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information. *Journal of Biomedical Informatics*, **40**(1):30-43.

[25] Yang B, Sayers S, Xiang Z, He Y (2011). Protegen: a web-based protective antigen database and analysis system. *Nucleic Acid Research*, **39**(Database issue):D1073-8.

[26] Frank G, Wheaton B, Bakalov V *et al.* (2009). An Ontology for Designing Models of Epidemics Role of the Ontology in Building Models. *Proceedings of ICBO 2009*, 47-50.

[27] Lynch CO, Cunni C, Schripsema E *et al.* (2007). A Biosurveillance Platform for BioSense Message Analysis Using Integrated Reference Ontologies and Intelligent Agents. *2007 AAAI Fall Symposium*, 86.

[28] Xu H, Lu Y, Jiang M *et al.* (2010). Mining Biomedical Literature for Terms related to Epidemiologic Exposures. *2010 AMIA Annual Symposium Proceedings*, 897.

[29] Geller J, Morrey CP, Xu J *et al.* (2009). Comparing Inconsistent Relationship Configurations Indicating UMLS Errors. *2009 AMIA Annual Symposium Proceedings*, 193.

[30] Ferreira JD, Batista DS, Couto FM, Silva MJ (2010). The Geo-Net-PT/Yahoo! GeoPlanet (TM) concordance. Technical Report #2010;5, Departamento de Informática da Faculdade de Ciências da Universidade de Lisboa. doi:10455/6677.

[31] Demographic and Social Statistics *in* United Nations Statistics Division. http://unstats.un.org/unsd/demographic/default.htm Accessed on 2012-01-05.

[32] Hitzler P, van Harmelen F (2010). A reasonable semantic web. *Semantic Web*, **1**(1):39-44.

[33] Couto FM, Silva MJ, Coutinho PM (2005). Semantic similarity over the Gene Ontology: family correlation and selecting disjunctive ancestors. *Proceedings of the 14th ACM,* 343-344.

[34] Pesquita C, Faria D, Bastos H, Ferreira AEN, Falcão AO, Couto FM (2008). Metrics for GO based protein semantic similarity: a systematic evaluation. *BMC Bioinformatics*, 9 Suppl 5, S4.

[35] Ferreira JD, Couto FM (2010). Semantic Similarity for Automatic Classification of Chemical Compounds. *PLoS Computational Biology*, **6**(9): e1000937. doi:10.1371/journal.pcbi.1000937.

[36] Cruz IF, Stroe C, Caimi F *et al.* (2011). Using AgreementMaker to Align Ontologies for OAEI 2011 . Ontology Matching Workshop at the 10th International Semantic Web Conference 2011.

[37] Gruber T (2008). Encyclopedia of Database Systems, chap. Ontology. Springer-Verlag. 9.

[38] Pesquita C, Faria D, Falcão AO, Lord PW, Couto FM (2009). Semantic similarity in biomedical ontologies. *PLoS Computational Biology*, 5(7): e1000443. doi:10.1371/journal.pcbi.1000443.

[39] Resnik P (1999). Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. *Journal of Art Intel Research*, **11**:95-130.

[40] Euzenat J, Shvaiko P (2007). *Ontology matching*. Springer-Verlag. New York Inc.

[41] Melnik S, Garcia-Molina H, Rahm E (2001). Similarity flooding: a versatile graph matching algorithm and its application to schema matching. *18th Int Conf on Data Engineering*, 117-128.

[42] Cruz IF, Sunna W (2008). Structural Alignment Methods with Applications to Geospatial Ontologies. *Transactions in GIS*, **12**:683-711.

[43] Hirschman L, Park J, Tsujii J *et al.* (2002). Accomplishments and challenges in literature data mining for biology. *Bioinformatics*, **18**(12):1553-1561.

[44] Rebholz-Schuhmann D, Kirsch H, Couto F (2005). Facts from text - is text mining ready to deliver?. *PLoS Biology*, **3**(2):e65. doi:10.1371/journal.pbio.0030065.

[45] Zweigenbaum P, Demner-Fushman D, Yu H, Cohen KB (2007) Frontiers of biomedical text mining: current progress. *Briefings in Bioinformatics*, **8**(5):358.

[46] Pinto HS, Tempich C , Staab S (2009). *Handbook on Ontologies*, "Ontology Engineering and Evolution in a Distributed World Using DILIGENT". Springer Berlin Heidelberg, Berlin, Heidelberg.

[47] Agirre E, Ansa O, Hovy E, Martinez D (2000). Enriching very large ontologies using the WWW. *Proceedings of the ECAI Ontology Learning Workshop*.

[48] Velardi P, Fabriani P, Missikoff M (2001). Using text processing techniques to automatically enrich a domain ontology. *Proc of the ACM Int Conf on Formal Ontology in Information Systems*. **34**:39.

[49] Ruiz-Casado M, Alfonseca E, Castells P (2005). Using context- window overlapping in synonym discovery and ontology extension. *Proceedings of RANLP-2005,* **36**:39.

[50] Witschel H (2005). Using decision trees and text mining techniques for extending taxonomies. *Learning and Extending Lexical Ontologies by using Machine Learning Methods*, **61**.